An Economic Approach to Gross Error Detection in Instrument Networks

Rollin Brant

University of Minnesota
School of Statistics
Technical Report #481

October 1986

University of Minnesota
School of Statistics
Department of Applied Statistics
St. Paul, Minnesota 55108

AN ECONOMIC APPROACH TO GROSS ERROR DETECTION IN INSTRUMENT NETWORKS

Rollin Brant
Department of Applied Statistics
University of Minnesota

October 1986

An Economic Approach to Gross Error Detection in Instrument Networks

Rollin Brant
Department of Applied Statistics
University of Minnesota
St. Paul, MN 55108

This article describes approaches to the detection and identification of gross errors occurring in instrument networks described by linear models. By adopting expected costs as the fundamental criterion and incorporating information concerning the reliability of specific components, cost-effective methods result. In addition, many of the logical difficulties associated with the application of previously developed "outlier" tests are avoided. Three basic models are considered and compared.

Key Words: Outlier detection; Linear models; Economic optimality; Inspection policies; Replacement policies.

AN ECONOMIC APPROACH TO GROSS ERROR DETECTION IN INSTRUMENT NETWORKS

1. Introduction

The monitoring and control of certain industrial processes depends on measurements from networks of instrumentation whose joint behaviour can be described by the linear model. In this article, methods for detecting the occurrence of gross errors and identifying their sources are considered. While apparently relevant methods, such as outlier tests, have already received extensive treatment, there is considerable room for new development. In particular, it is desirable to take into account such fundamental objectives as economically efficient operation. Additionally, pertinent information available to process operators, such as reliablity data, should be incorporated when available. Taking these factors into consideration yields significant improvements in cost effectiveness over the commonly used approaches.

For the sake of concreteness, we shall focus on chemical process networks, which are described briefly below, and at length in Tamhane and Mah (1985). A chemical process network consists of a set of process units, interconnected by a network of streams, through which various materials flow. Instruments are attached at various points in the network, yielding mass flow rates, concentrations, and a variety of measurements which are subject to error. Owing to mass and energy balance constraints, the true values of these measurements satisfy certain linear restrictions. In addition, certain measurements of interest cannot be obtained directly, but must be inferred from the available measurements through use of the balance constraints. Overall,

it is possible to re-express the unknowns in terms of some linear regression model, with a weighted and possibly dependent error structure.

Two aspects of the gross error problem must be distinguished: 1. detection and identification of gross errors and, 2. data reconciliation, by which is meant the problem of making reliable inferences about the true values. It is clear that the problems are closely linked, in particular, that some allowance for the possibility of gross errors in the observed data must be made at the data reconciliation stage. It must however be recognized that detection and identification is a problem that is logically distinct from the latter. If for instance, gross errors provide evidence of malfunction, either in instruments or in other components of the network, then the implications of detection extend beyond the immediate data reconciliation.

Recognizing that the two problems require separate consideration, our attention here is directed to the detection problem. We begin in Section 2 with a brief review and critique of current methodology. Section 3 follows with a discussion of modelling considerations in the present context. Sections 4, 5 and 6 describe various models, and the associated methods. Section 7 compares the various approaches.

2. Background

We shall assume, following Tamhane and Mah (1985), that one has available a vector, $\mathbf{y}:n\times 1$, of measurements, with a corresponding vector of unknown true values $\boldsymbol{\eta}:n\times 1$. In addition, we assume a set of values corresponding to unmeasurable variables is described by the vector $\boldsymbol{\xi}:m\times 1$, which is linearly related to $\boldsymbol{\eta}$ through a set of constraints

$$A\boldsymbol{\xi}+B\boldsymbol{\eta}=\mathbf{c}, \qquad (2.1)$$

where $A: q \times m$, and $B: q \times n$ and $\mathbf{c}: q \times 1$ are known quantities. Letting $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\eta}$ represent the measurement errors, it is further assumed that, under nominal conditions, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and that $Var(\boldsymbol{\epsilon})$ is known. Additionally, it is conventional to assume that $\boldsymbol{\epsilon}$ is distributed as a Gaussian variate.

One can reduce the above to a standard linear model by using (2.1) to rewrite $\boldsymbol{\eta} = X\boldsymbol{\theta}$ and $\boldsymbol{\xi} = Z\boldsymbol{\theta}$, where X and Z are known, as shown by the following.

Proposition 1: Let $A: q \times m$, $B: q \times n$ and $\mathbf{c}: q \times 1$ be given. There exist matrices X and Z, such that, for all $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfying $A\boldsymbol{\xi} + B\boldsymbol{\eta} = \mathbf{c}$, one can write $\boldsymbol{\eta} = X\boldsymbol{\theta}$ and $\boldsymbol{\xi} = Z\boldsymbol{\theta}$, for some $\boldsymbol{\theta}$.

Proof: see appendix.

By the above proposition, the model can be re-expressed in the form $\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$, so that standard least squares theory suffices to derive estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, and more or less standard diagnostic methodology can be applied. Proposed methods considered so far have been based on examination of the residuals, $\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\eta}}$ denotes the standard (weighted) least squares estimate of $\boldsymbol{\eta}$. More explicitly, the data is first examined for outliers using outlier tests, "significant" outliers are taken as representing gross errors, and the remaining data is then analyzed using standard regression methods.

A variety of proposals of this form are reviewed in Tamhame and Mah (1985). In particular, the authors consider methods based on the use of $\mathbf{e}^* = \Sigma^{-1}\mathbf{e}$ for detecting individually discrepant observations. Tamhane

(1982) proposed the rule

$$|\{var(e^*_i)\}^{-1/2} e^*_i| > k,$$

for deciding on the presence of a gross error in the i'th observation, where k is the upper $1/2 \cdot \{1-(1-\alpha)^{1/n}\}$ point of the standard Gaussian distribution. Subsequently data reconciliation then consists of analyzing the data which passes the above test.

This common sense strategy is susceptible to improvement, mainly on the grounds that it attempts to solve two distinct problems with one technique, namely, outlier testing. Though it is natural to characterize gross error detection as being basically an outlier identification problem, data reconciliation is primarily an estimation/prediction problem. General arguments for seeking distinct solutions to such problems are offered by Barnett and Lewis (1984). Practical support for this view is provided by the experimental work of Ruppert and Carroll (1980), who show how poorly outlier testing schemes perform when assessed in terms of the performance of the subsequent estimates. It is thus essential that the problems be given separate consideration.

Further potential for improvements emerge when one considers the nature of outlier testing itself. The conventional framework adopted for the development of outlier tests does not adequately reflect the particular aims of the given situation. Outlier tests are general purpose techniques which tend to be based on loose foundations. Such imprecision is an unavoidable consequence of the fact that the aims of outlier detection vary considerably from context to context, coupled with fact that a variety of mechanisms for generating outliers are possible. Owing to the more specialized context here, it is only natural to seek appropriately constructed methods, rather than apply the all-purpose

tools represented by conventional outlier tests.

3. Models for Gross Error Detection

Implicit in the aims of detection and identification, as distinct from data reconciliation, is the notion that a occurence of a gross error can be identified with some identifiable, and practically significant event, such as a malfunction, conceivably having ramifications beyond the particular data at hand. Consequently we imagine that the imputation of such an error's occurrence leads to certain well-defined actions, e.g. inspection or replacement of certain components of the network. Since our ultimate aim is the economically efficient operation of the process, the consideration of the attendent costs of such actions will be a key component in the decision process. Other features of the networks, such as reliability of components are also relevant.

A natural place to begin is to augment the model put forward in section 2 to account for these particular features. For the sake of utmost generality, we shall consider that there is a well defined set of r potential gross errors, whose joint occurrence or non-occurrence shall be recorded by a vector $\zeta=(\zeta_1,\zeta_2,\ldots,\zeta_r)$, with $\zeta_j=1$ if error j has occurred and $\zeta_j=0$ otherwise. The aim of our detection procedures is to impute some value to ζ, which we will denote $\hat{\zeta}$:r×1. In many cases it will be natural to take r=n, and identify gross errors with particular instruments and the resultant measurements. The larger framework allows such events to affect the network more generally.

The Gaussian assumption of the nominal model of Section 2 can be taken as specifying the conditional distribution of ϵ given $\zeta=\mathbf{0}$. A comprehensive model for the behaviour of the system can be based on

describing the behaviour of ϵ given all possible configurations of ζ. Often it will be convenient to model this behaviour in terms of shifts in mean and variance of the distribution of ϵ, as is considered in the Section 4. In Section 5 a more computationally convenient Gaussian model is considered. In Section 6 a slightly more pragmatic view of the characterization of gross errors is taken, which also leads to straightforward calculations.

While the considerations above bear some resemblance to previously considered models for outlier identification (see e.g. Cook and Weisberg 1982) an important distinction is that here the consequences of taking $\hat{\zeta}_j=1$ are taken to be well defined, i.e. lead to predetermined remedial actions, such as inspection and/or replacement of system components. Naturally associated with the given actions will be certain costs. For a given malfunction, with indicator, ζ, the cost which attends taking $\hat{\zeta}=k$ when $\zeta=l$ is assumed to specified by c_{kl}. Additionally, we assume that overall costs are additive over the r errors to be considered.

The specification of these costs clearly hinge on the type of action occasioned by taking $\hat{\zeta}=1$. For example take the case that $\zeta=1$ corresponds to a given malfunction and $\hat{\zeta}=1$ corresponds to the decision to inspect, and repair if necessary, the component in question. Generally speaking, c_{00} can be set to 0. c_{01} represents the cost of letting a malfunction go undetected, which we denote c_M. The costs c_{10} and c_{11} will both include an inspection cost, c_I and in the case of $\zeta=1$, a repair cost, c_R. Thus, the cost structure for an inspection policy is $c_{00}=0$, $c_{01}=c_M$, $c_{10}=c_I$, and $c_{11}=c_I+c_R$. On the other hand, if an automatic component replacement policy is applied, the structure would take the

form $c_{00}=0$, $c_{01}=c_M$, $c_{10}=c_{11}=c_R$. By considering such costs, one can take an economic view and adopt the most directly relevant criteria for the formulation of methodology. For example, see Lorenzen and Vance (1986), for the implications of such a view with regard to the design of control charts.

Aside from economic considerations, another important feature which distinguishes the present situation from the usual regression framework, is that the observations arise from specific types of components of, generally speaking, standard manufacture. It is natural then to consider the reliability of these components as being characterized by a probability distribution $p(\zeta)$ for ζ. Here we assume that sufficient technical knowledge or prior experience exists to make this specificiation explicitly. Failing this it will be necessary to estimate $p(\zeta)$ from monitoring the operation of the process, a problem which requires further investigation, not undertaken here.

The type of model described above suggests a fully Bayesian approach, in that prior probabilities for gross errors are assumed. The model, however, falls short of a full Bayesian framework in not assigning priors for the remaining parameters in the model. Thus a complete Bayesian solution to the decision problem is not considered. To circumvent difficulties associated with the unknown parameters, a not always efficient but practical approach is to base decisions on the residuals, $\mathbf{e}$, derived from an appropriate model, for instance, the nominally held model with $\zeta=\mathbf{0}$. Since $\mathbf{e}$ is a function of ϵ, assuming the correct specification of the model, computation of the conditional distribution of ζ given $\mathbf{e}$, $p(\zeta \mid \mathbf{e})$, is feasible whenever a complete specification along the lines indicated above can be made.

Combining $p(\zeta \mid \mathbf{e})$ and the cost structure will give rise to Bayesian decision rules in straightforward manner. Under the assumption of additive (over errors) costs, expected costs overall can be minimized by considering individual errors, ζ, choosing $\hat{\zeta}=1$ whenever

$$\Pr\{\zeta=1 \mid \mathbf{e}\} > c,$$

where $c=(c_{01}-c_{00})/\{(c_{10}-c_{00})+(c_{01}-c_{11})\}$ is a cost ratio. In the case of an inspection policy, $c=c_I/(c_M-c_R)$, i.e. the ratio of inspection costs to the potential savings due to replacing a defective component. When automatic replacement is the rule, $c=c_R/c_M$.

The simple form of the rule above belies the practical difficulties of implementation in practice. The chief problems to be overcome are in making appropriate specifications of the above parameters, and in completing the required probability calculations. We now consider some possible approaches.

4. The Contaminated Gaussian Model

One commonly used model for the generation of gross errors is the Gaussian contamination model. In this context we are lead to adopt such a model by assuming that gross errors can be characterized in terms of corresponding additive disturbances. Specifically, we again assume that $\mathbf{y}=\boldsymbol{\eta}+\boldsymbol{\epsilon}$, but suppose $\boldsymbol{\epsilon}$ can be written as $\boldsymbol{\epsilon}=\boldsymbol{\epsilon}_0+\sum\zeta_i\boldsymbol{\tau}_i$, where $\boldsymbol{\epsilon}_0$ is a Gaussian variate with $\mathbf{0}$ mean and known variance Σ and the $\boldsymbol{\tau}_i$'s are independently distributed, Gaussian variates with means $\boldsymbol{\delta}_i$ and variance-covariance matrices, Ψ_i, which are assumed known in the following. According to the particular choice for $p(\zeta)$, the resulting marginal distribution for $\mathbf{y}$ is a mixture of normals, i.e. a contaminated

Gaussian distribution.

The most common model of this type assumes there are r=n potential gross errors, each associated with a particular observation. In particular, the two most practically useful instances of the above model are the mean shift and variance inflation models. Letting $\mathbf{u}_i$ denote a p-vector of 0's except for 1 in the i'th place, the mean shift model assumes that $\delta_i=\delta_i\mathbf{u}_i$, and that the Ψ_i's are all 0 matrices, i=1,...,n. The variance inflation model assumes the δ_i's are 0 vectors, and that $\Psi_i=\psi_i^2\mathbf{u}_i\mathbf{u}_i^t$.

As argued above, a natural strategy is to base inference on the conditional distribution of ζ given the residual vector, **e**, most conveniently calculated in the form,

$$p(\zeta_0\,|\,\mathbf{e})=\{f(\mathbf{e}\,|\,\zeta_0)\cdot p(\zeta_0)\}\,/\,\{\textstyle\sum_\zeta f(\mathbf{e}\,|\,\zeta)\cdot p(\zeta)\}, \qquad (4.1)$$

where $f(\mathbf{e}\,|\,\zeta)$ denotes the conditional density of **e** given ζ. As noted earlier, assuming additive costs, optimal decisions concerning ζ_j are based on the associated cost ratio, c_j, and $P\{\zeta_j=1\,|\,\mathbf{e}\}$, which can be determined straightforwardly from the above.

Two computational issues affect the application of this apparently elementary procedure. The first problem is essentially superficial, arising out the fact that the distribution of **e** is singular, complicating the evaluation of the required conditional densities in the above. The following proposition is useful in achieving a convenient representation.

Proposition 2: Suppose that **x** is multivariate Gaussian, with

mean $\mathbf{m}$ and variance $I+M$, where M is positive semi-definite. If H is an idempotent matrix, of rank p, then the density of $\mathbf{r}=(I-H)\mathbf{x}$ is proportional to

$$\{|I+M|\,|Q^t(I+M)^{-1}Q|\}^{-1/2}\exp(-1/2c),$$

where $c=\mathbf{d}^t[(I+M)-Q\{Q^t(I+M)^{-1}Q\}^{-1}Q^t]\mathbf{d}$,

$\mathbf{d}=(I+M)^{-1}(\mathbf{r}-(I-H)\mathbf{m})$, and

$Q:n\times p$ is a matrix whose columns are the eigenvectors of H.

Proof: See appendix

To apply the result in the given context, we first suppose that $\Sigma_0=KK^t$, for some invertible $K:n\times n$ and let $\tilde{\mathbf{y}}=K^{-1}\mathbf{y}$ and $\tilde{X}=K^{-1}X$. The residual $\mathbf{e}$, from the weighted least squares fit relative to Σ_0 can be written as $K\tilde{\mathbf{e}}$, where $\tilde{\mathbf{e}}$ is the unweighted residual from the regression of $\tilde{\mathbf{y}}$ on $\tilde{X}=(I-H)\tilde{\mathbf{y}}$. Now the sought after conditional density of $\mathbf{e}$ given ζ is proportional to that of $\tilde{\mathbf{e}}$ given ζ. Proposition 2 applies to yield this density by taking $\mathbf{x}=\tilde{\mathbf{y}}$, $\mathbf{m}=\tilde{X}\boldsymbol{\theta}$, $M=K^{-1}(\sum\zeta_j\Psi_j)(K^{-1})^t$ and $H=\tilde{X}(\tilde{X}^t\tilde{X})\tilde{X}^t$. Noting that if X is of full rank, then $\tilde{X}=QR$, for some invertible matrix R, and letting $\boldsymbol{\delta}_\zeta=\sum\zeta_j\boldsymbol{\delta}_j$ and $\Psi_\zeta=\sum\zeta_j\Psi_j$, we have that $f(\mathbf{e}\,|\,\zeta)$ is proportional, as a function of $\mathbf{e}$ and ζ, to

$$\{|\Sigma+\Psi_\zeta|\cdot|X^t(\Sigma+\Psi_\zeta)^{-1}X|\}^{-1/2}\cdot$$

$$\times\exp(\mathbf{f}_\zeta{}^t[\Sigma+\Psi_\zeta-X\{X^t(\Sigma+\Psi_\zeta)^{-1}X\}^{-1}X^t]\mathbf{f}_\zeta),$$

where $\mathbf{f}_\zeta=(\Sigma+\Psi_\zeta)^{-1}\{\Sigma-X(X^t\Sigma^{-1}X)^{-1}X^t\}\Sigma^{-1}(\mathbf{e}-\boldsymbol{\delta}_\zeta)$.

A more significant potential stumbling block to the routine computation of $P\{\zeta_j=1\,|\,\mathbf{e}\}$ arises, for example, when it is assumed that

$p(\zeta)=\prod p_j^{\zeta_j}$, i.e. that gross errors occur independently with probabilities, p_j, j=1,..,r. The result of this is a combinatorial explosion in the number of terms in the denominator of (4.1). When r is large an ad hoc but sensible approach is to neglect improbable multiple errors in the computation. This is equivalent to conditioning on the occurrence of a limited number of errors, and serves to reduce, but not eliminate the potential computational burden. Thus the present approach is practically feasible mainly when r is of moderate size.

Just as important as computational problems are the difficulties associated with completing the specification of the model given above. As discussed previously, the chief practical difficulty is likely to be specifying the δ's and Ψ's for the contaminating disturbances. In the ideal, these parameters should be specified on the basis of prior data concerning the behaviour of components under errors. Practically speaking, such information may be difficult to obtain, and parameter choice must be made on a subjective basis.

Characterization of gross errors in terms of shifts in means and/or varianced may be somewhat unnatural to field workers. A mathematically equivalent, but perhaps more transparent approach to this specification arises from considering hypothetically that true values, η, are known, so that measurement errors, ϵ, can be determined. In this instance, decisions about the integrity of components would presumably be based directly on ϵ, leading one to consider $p(\zeta \mid \epsilon)$. In some settings this function may provide a more natural basis for completing the specification required above, as in the variance inflation model.

Assuming that Σ is diagonal with entries, σ_i^2, i=1,...,n, and gross errors occur independently, the above approach leads one to consider the logistic form,

$$Pr\{\zeta_i=1 \mid \epsilon_i\} = \exp(\alpha_i)/\{1. + \exp(\alpha_i)\}^{-1},$$

where $\alpha_i = \log\{p_i\sigma_i/(1-p_i)\omega_i\} + 1/2(\sigma_i^{-2} - \omega_i^{-2})\cdot\epsilon_i^2$ and $\omega_i^2=\sigma_i^2+\psi_i^2$. If one can specify a critical error value, ϵ_c, which signals an gross error with given probability, p_c, then ψ_i^2 can be determined by setting $\alpha_i=\log\{p_c/(1-p_c)\}$ and solving for ψ_i.

5. The Pure Gaussian Model

The assumption of Gaussian errors in applications of the linear model is in a large sense one of convenience. As in many modelling decisions, this choice is based mainly on considerations of tractability, since more objective criteria are generally too hard to assess. In the above section, the use of Gaussian contaminants was similarly an arbitrary choice, dictated largely by convention, and somewhat by tractability, though as we have seen, the results provided by the model can lead to computational difficulties. In this section we avoid such difficulties by assuming that the mixing of routine and gross errors results in a Gaussian distribution.

The model we shall adopt accounts for the possibility of error j's occurrence in terms of bias components, γ_j and increases in variability, Λ_j, in the Gaussian distribution of the error term, ϵ. Unconditionally (i.e. in the absence any knowledge concerning ζ) ϵ is Gaussian, with mean $\gamma=\sum\gamma_j$ and variance-covariance matrix $\Omega=\Sigma+\sum\Lambda_j$, whereas

conditional on the absence of all gross errors, ϵ is Gaussian, with mean **0** and variance Σ. The specification of the model is completed by considering the behaviour of the system conditional on the absence of specified error. More precisely, let $G=\{j \mid \zeta_j=0\}$ and let S be a subset of $\{1,2,\ldots, r\}$. We assume that the conditional distribution of ϵ given $S \subseteq G$ is Gaussian with mean $\boldsymbol{\gamma}-\sum_{j\epsilon S}\boldsymbol{\gamma}_j$ and variance $\Omega-\sum_{j\epsilon S}\Lambda_j$. Certain restrictions on $p(\boldsymbol{\zeta})$, the $\boldsymbol{\gamma}$'s, and Λ's, must hold for this distribution to be well defined, but given these restrictions, the components of the implicitly defined mixture distribution for ϵ can be determined.

This model has several appealing features. To illustrate the first of these, we consider the variance inflation form of the model, described by taking r=n, with $\boldsymbol{\gamma}_i=\mathbf{0}$ and $\Lambda_i=\lambda_i^2\mathbf{u}_i\mathbf{u}_i^t$. Under these assumptions, the conditional density of an error, ϵ_i, given that it stems from a gross error is given by

$$f(\epsilon_i)=p_i^{-1}(2\pi)^{-1/2}\{\omega_i^{-1}\exp(-\epsilon_i^2/2\omega_i^2)-(1-p_i)\sigma_i^{-1}\exp(-\epsilon_i^2/2\sigma_i^2)\},$$

where $\omega_i^2=\sigma_i^2+\lambda_i^2$. This "contaminating" density is well defined only if $(1-p_i)\cdot(\omega_i/\sigma_i)\leq 1$, reflecting that the constraints on $p(\boldsymbol{\zeta})$, the $\boldsymbol{\gamma}_j$'s and Λ_j's mentioned above. Interestingly, it is density is bimodal for $(1-p_i)\cdot(\omega_i/\sigma_i)^3 > 1$. Figure 1 illustrates a typical form, illustrating that the model places contaminants where one would expect them to be, i.e. in the tails of the "good" distribution, naturally reflecting the abberant behaviour one associates with a malfunctioning instrument.

The second advantage which stems from the model is its initial motivation, i.e. computational simplicity. The general approach given in Section 3 based on the residual vector applies, so that once again

decision rules can be based on $\Pr\{\zeta_j=1 \mid \mathbf{e}\}$. Under the present assumptions, it is natural to take the residual, $\mathbf{e}$, arising from the weighted fit relative to Ω and adjusted for any known bias, $\boldsymbol{\gamma}$. This is most conveniently expressed as $\mathbf{e}=K\tilde{\mathbf{e}}$, where $\tilde{\mathbf{e}}$ is the residual from the unweighted regression of $\tilde{\mathbf{y}}=K^{-1}(\mathbf{y}-\boldsymbol{\gamma})$ on $\tilde{X}=K^{-1}X$, K being chosen such that $\Omega=KK^t$.

In the general case, $\Pr\{\zeta_j=0 \mid \mathbf{e}\}$ is straightforwardly given by $\{f(\mathbf{e} \mid \zeta_j=0)(1-p_j)\}/f(\mathbf{e})$, where $f(\mathbf{e} \mid \zeta_j=0)$ and $f(\mathbf{e})$ denote conditional and unconditional densities for $\mathbf{e}$. Here the relevant distributions are singular multivariate Gaussians, but the same arguments used in Section 4 apply to facilitate evaluation of the densities, yielding that $f(\mathbf{e})$ is proportional to $\{|\Omega| \cdot |X^t\Omega^{-1}X|\}^{-1/2} \exp(-1/2\mathbf{e}^t\Omega^{-1}\mathbf{e})$, and that $f(\mathbf{e} \mid \zeta_j=0)$ is proportional to

$$\{|\Omega-\Lambda_j| \cdot |X^t(\Omega-\Lambda_j)^{-1}X|\}^{-1/2}$$
$$\times \exp(\mathbf{f}_j{}^t[\Omega-\Lambda_j-X\{X^t(\Omega-\Lambda_j)^{-1}X\}^{-1}X^t]\mathbf{f}_j),$$

where $\mathbf{f}_j=(\Omega-\Lambda_j)^{-1}\{\Omega-X(X^t\Omega^{-1}X)^{-1}X^t\}\Omega^{-1}(\mathbf{e}+\boldsymbol{\gamma}_j)$.

In the variance inflation form of the model, if Σ is diagonal with entries, σ_i^2, and gross errors are independent, $\Pr\{\zeta_i=0 \mid \mathbf{e}\}$ depends only on e_i and is given by

$$(1-p_i)\{(\sigma_i^2+\lambda_i^2)/(\sigma_i^2+\lambda_i^2h_{ii})\}^{-1}$$
$$\times \exp[-(1/2)\lambda_i^2e_i^2\{(\sigma_i^2+\lambda_i^2)\cdot(\sigma_i^2+\lambda_i^2h_{ii})\}^{-1},$$

where h_{ii} is the i'th diagonal element of $H=\tilde{X}(\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t$.

As in the model considered in Section 4, difficulty in applying this model may arise in the specification of the $\boldsymbol{\gamma}$'s and Λ's. In situations

where background information is sparse, it will reasonable to adopt a simple model, such as the variance inflation model for gross errors. Owing to the bounds placed on λ_i^2 by the mixture formulation, we require that $(\lambda_i^2/\sigma_i^2) \leq (1-p_i)^{-2}-1$. Once more consideration of $Pr\{\zeta_i=1 \mid \epsilon_i\}$ may be helpful. Letting $\omega_i^2=\sigma_i^2+v_i^2$, we have that

$$Pr\{\zeta_i=1 \mid \epsilon_i\} = 1-(1-p_i)(\omega_i/\sigma_i)\cdot\exp\{1/2(\omega_i^{-2}-\sigma_i^{-2})\epsilon_i^2\}$$

may lead to a choice for λ_i^2 if some tolerance level ϵ_c can be established with associated gross error probability p_c. Failing this, an ad hoc, but intuitively appealing choice is to take λ_i^2 at the upper limit of the bounds determined by the choice of p_i. In particular this implies that the conditional probability of a gross error having occurred given $\epsilon_i=0$ is 0.

6. A Pragmatic Approach to Modelling Gross Errors.

In the development so far, the characterization of gross errors has been model-based, in that a model for the behaviour of responses given the occurrence of gross errors is explicitly considered. A more pragmatic approach is to characterize the occurrence of a gross errors strictly in terms of the values of ϵ_i, i=1,..,n. For example, if gross errors can be associated with particular observations, a plausible characterization might be that $\zeta_i=1$, whenever $|\epsilon_i| > \kappa_i$ and 0 otherwise, reflecting the natural desire to detect any observation incorporating an error which exceeds given tolerances.

We assume once more that $y=\eta+\epsilon$, requiring here that ζ be given

directly as a function of ϵ, $\zeta=\zeta(\epsilon)=(\zeta_1(\epsilon),\ldots,\zeta_r(\epsilon))$. Again it is assumed that the distribution of $\epsilon=\mathbf{y}-\boldsymbol{\eta}$, allowing for gross errors, is multivariate Gaussian, with known mean, $\boldsymbol{\mu}$, and variance, Σ. In the absence of precise knowledge of the bias induced by gross errors it will usual to take $\boldsymbol{\mu}=\mathbf{0}$. Σ must be chosen to reflect any excess variance due to gross errors. This can be based on past operating records, or on reliability data for the given components.

Assuming additive costs, then, optimal decisions are based once again on $\Pr\{\zeta_j=1\,|\,\mathbf{e}\}=E\{\zeta_j(\epsilon)|\mathbf{e}\}$, in conjunction with the assumed cost ratio, c_j. Defining $\mathbf{e}$ as in section 5, the distribution of ϵ conditional on $\mathbf{e}$ is multivariate Gaussian, with mean $\mathbf{m}=\mathbf{e}+\boldsymbol{\mu}$ and variance-covariance matrix, $V=X(X^t\Sigma^{-1}X)^{-1}X^t$. Since $\zeta_j(\epsilon)$ is just an indicator function, $E\{\zeta_j(\epsilon)|\mathbf{e}\}$ is, in principle, determinable as the probability content of the corresponding region. Taking the simplest case based on critical tolerances, κ_i, for individual measurements, $\zeta_i(\epsilon)$ depends only on ϵ_i, whose conditional distribution given $\mathbf{e}$ is Gaussian with mean $m_i=e_i+\mu_i$ and variance v_{ii}, the i'th diagonal element of V. Thus,

$$\Pr\{\zeta_j=1\,|\,\mathbf{e}\} = \Phi\{(-\kappa_i-m_i)/\sqrt{v_{ii}}\} + [1 - \Phi\{(\kappa_i-m_i)/\sqrt{v_{ii}}\}].$$

Part of the appeal in the above strategy is that it simplifies to some extent the specification of "reliability" parameters. For example, the simplest useful instance of the model will be as given above, with the added restrictions that $\boldsymbol{\mu}=\mathbf{0}$, and Σ diagonal, in which case only σ_i^2 and κ_i, need be specified. Estimation of σ_i^2 on the basis of previous data may be less complicated than one might suspect, since it is likely

that such data, unless carefully screened, will incorporate the types of gross errors that one seeks to detect. Thus more or less standard estimation methods can be applied.

Specification of κ_i can be approached in a number of different ways. Manufacturer's specifications for reliability will in general give sensible guidelines as to maximum feasible errors under nominal conditions. Alternately, if some underlying model is entertained for characterizing ζ, κ_i can be chosen in such a way to maximize the correspondence between the event, $|\epsilon_i| > \kappa_i$ and ζ_i.

7. Comparing Error Detection schemes.

A natural starting point in evaluating the above proposals is a comparison with the outlier tests. In the context under consideration, outlier tests have been recommended for the detection of potentially suspect observations, and thus can be regarded as inspection rules. As suggested earlier, such tests tend to be inappropriate in the given setting, and we consider this in somewhat more detail here.

Outlier testing methods are invariably formulated under the framework of hypothesis tests. Various developments along these classical lines are reviewed in Beckman and Cook (1983) and Hawkins (1980). Fundamental to the such approaches is control over the Type I error rate. Since candidates for outliers are selected a posteriori, adjusting for multiplicity to maintain an "overall" type I error rate has generally been deemed advisable. In some places, adjustments based on the Bonferroni inequality are made to achieve this, or as in the method suggested by Tamhane (1982), appeal to the "near-independence" of residuals can be made to justify the use of an individual error rate

of $\{1-(1-\alpha)^{1/n}\}$.

From the economic viewpoint, this type of adjustment is generally counter-productive. If, for instance, one assumes that gross errors occur more or less independently of each other and that costs are additive, costs are best controlled by minimizing costs on a purely per item basis. The effect of Bonferroni adjustments, on the other hand, is to reduce the per item type I rate, while greatly inflating the per item type II error rate. If minimization of costs is an objective, and hypothesis tests are to be applied, the per item rate is the relevant quantity to control.

Lack of control over the type II error rate is in fact a fundamental drawback to the application of the hypothesis testing framework to the present situation, even when no adjustments for multiplicity are made. This is the result of the emphasis attached to maintaining specified significance levels in the classical testing framework. One convenient way to assess the possible ramifications of this mis-emphasis is to interpret traditional rules as economically based ones, with implicitly defined costs, etc.

For simplicity we consider the simplest and hypothetical case of a single observation, y, made with known η and ϵ, where the nominal assumption is that ϵ is Gaussian with mean 0 and variance σ^2. In this situation, a traditional rule would be to assess the observation as a gross error if $\epsilon^2 > CRIT_0 = (z_{\alpha/2})^2$. Consider by way of contrast a Gaussian mixture model that supposes that with probability p the variance is inflated to ψ^2, with cost ratio c. Under this approach the corresponding rule is based on

$$CRIT_M = 2(1-\sigma^2/\psi^2)^{-1} \cdot \log\{(\psi/\sigma)(p^{-1}-1)/(c^{-1}-1)\}.$$

By equating the two critical values, one can assess the traditional rule from an economic standpoint. Setting $CRIT_0 = CRIT_M$ implies that if the ratio of the nominal variance to that for gross errors is υ^2, then c and p satisfy

$$(c^{-1}-1)/(p^{-1}-1) = h(\upsilon,\alpha) = \upsilon \exp\{-1/2\cdot(1-\upsilon^{-2})(z_{\alpha/2})^2\}.$$

Contours of $h(\upsilon,\alpha)$ are plotted in Figure 2 and in Figure 3, $h(\upsilon,\alpha)$ is plotted versus υ for α fixed at .05. The latter plot reveals that the classical 5% rule implicitly assumes that an upper bound on c is, roughly speaking, p/2. Thus, though the classical rule does not explicitly take costs into account, it tends to the implicit assumption that the ratio c is small. Generally speaking this will be reasonable, for instance, under an inspection policy. In this case $c=c_I/c_M$, and since costs associated with inspections will generally be low compared to the costs which derive from malfunctioning instruments, small values of c will be more common than not.

With regard to comparisons of the methods proposed in sections 4, 5, and 6, since they arise out of different models for the gross error process, the usual criteria, such as power or expected costs, are not strictly relevant. The ideal criterion for choosing between the three, then, is correspondence with the actual behaviour of the process to be modelled. Three pragmatic criteria, parsimony, flexibility, and tractability, are often, however, more important.

The Gaussian contamination model offers a great deal of flexibility, but at the cost of incorporating a large number of parameters through $p(\zeta)$, the δ_i's and Ψ_i's. As well, fairly extensive computations are involved in arriving at the necessary conditional probabilities, and, in

particular, care must be taken in avoiding combinatorial explosion in the number of terms evaluated. The utility of this model may well be restricted to fairly small networks, where the potential sources of gross error are not numerous.

The pure Gaussian model is attractive in that the necessary computations are not too complex. However, the model is somewhat restricted in its implications by the implicit bounds placed on the parameters. As well, quite a large number of parameters are involved in the specification of the model, which may impede its practical application.

The pragmatic model for gross errors has a great deal to recommend it. Generally speaking, it is the easiest to apply in practise, since its parameters are the simplest to specify and the attendant calculations are straightforward, at least in the case of independent errors. In addition, the underlying rationale in terms of characterizing gross errors is very pragmatic and easily understood.

6. Conclusion

Deciding what to do about outliers in linear models is a question that plagues data analysts. Broad recommendations are very risky to make, owing to the diversity of mechanisms giving rise to outliers. In addition the specific aims of model fitting, be they simple description, or more formal inferences, either with respect to particular parameters or to predictions, need to be taken into account. In situations, however, where aims can be clearly specified and mechanisms giving rise to outliers can be characterized, specific measures can be prescribed.

The problem of gross error detection in process networks affords is sufficiently well defined that the above questions can be sensibly

addressed. Familiar methods, such as conventional outlier tests, can then be assessed according to the most relevant critieria. Additionally, depending on the level of background information, "optimal" methods can be developed by taking such criteria into account at the initial stages of formulating methodology.

Much remains to be done in the development of economically based error detection techniques. In particular, estimation based on operating records, for parameters describing reliabilities and gross error behaviour needs to be considered. In addition, the serial nature of data taken on networks should be taken into account. By doing so, even greater improvements over ad hoc outlier testing are possible.

APPENDIX

Proposition 1: Let $A:q\times m$, $B:q\times n$ and $\mathbf{c}:q\times 1$ be given. There exist matrices X and Z, such that, for all $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfying

$$A\boldsymbol{\xi}+B\boldsymbol{\eta}=\mathbf{c}, \qquad (A.1)$$

one can write $\boldsymbol{\eta}=X\boldsymbol{\theta}$ and $\boldsymbol{\xi}=Z\boldsymbol{\theta}$, for some $\boldsymbol{\theta}$.

Proof: Without loss of generality we can assume that $\mathbf{c}=\mathbf{0}$, for by choosing any pair $(\boldsymbol{\xi}_0,\boldsymbol{\eta}_0)$ satisfying $A\boldsymbol{\xi}+B\boldsymbol{\eta}=\mathbf{c}$ one can simply work in terms of $\mathbf{y}'=\mathbf{y}-\boldsymbol{\eta}_0$, $\boldsymbol{\eta}'=\boldsymbol{\eta}-\boldsymbol{\eta}_0$, and $\boldsymbol{\xi}'=\boldsymbol{\xi}-\boldsymbol{\xi}_0$, which satisfy the above model with $\mathbf{c}=\mathbf{0}$. Letting $P:(q\times[q-s])$ be such that the column space of P spans the orthogonal complement of the s-dimensional column space of A ($s<q$ is necessary for (A.1) to represent any constraints on $\boldsymbol{\eta}$), we have that by premultiplying (A.1) by P^t and letting $M=P^tB$, that $M\boldsymbol{\eta}=\mathbf{0}$ is necessary and sufficient for $A\boldsymbol{\xi}+B\boldsymbol{\eta}=\mathbf{0}$, to hold for some $\boldsymbol{\xi}$. Supposing that M has rank p, and re-arranging the ordering of measurements $\mathbf{y}$, if necessary, we can write $M=[M_1 \mid M_1M_2]$, where M_2 is $([q-s]\times[n-p])$ and $M_1:([q-s]\times p)$ is of full rank, which by the previous implies that $\boldsymbol{\eta}_1+M_2\boldsymbol{\eta}_2=\mathbf{0}$, where $\boldsymbol{\eta}=(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)^t$. Thus, by letting $\boldsymbol{\theta}=\boldsymbol{\eta}_2$, $X=(-M_2^t \mid I)^t$, we have $\boldsymbol{\eta}=X\boldsymbol{\theta}$. It is easy to confirm that $P^tBX=0$, which implies that $BX=AZ$ for some $Z:(m\times p)$, so that $A\boldsymbol{\xi}+AZ\boldsymbol{\theta}=\mathbf{0}$. Assuming that A is of full column rank, this implies that $\boldsymbol{\xi}=Z\boldsymbol{\theta}$, i.e. is completely specified in terms of $\boldsymbol{\theta}$.

Proposition 2: Suppose that $\mathbf{x}$ is multivariate Gaussian, with mean $\mathbf{m}$ and variance I+M, where M is positive semi-definite. If H is an idempotent matrix, of rank p, then the density of $\mathbf{r}=(I-H)\mathbf{x}$ is proportional to

$$\{|I+M|\,|Q^t(I+M)^{-1}Q|\}^{-1/2}\exp(-1/2c),$$

where $c=\mathbf{d}^t[(I+M)-Q\{Q^t(I+M)^{-1}Q\}^{-1}Q^t]\mathbf{d}$,

$\mathbf{d}=(I+M)^{-1}(\mathbf{r}-(I-H)\mathbf{m})$, and

Q:n×p is a matrix whose columns are the eigenvectors of H.

Proof: We begin by noting that the distribution of $\mathbf{r}$ is multivariate Gaussian, with mean vector $(I-H)\mathbf{m}$, and variance-covariance matrix, $(I-H)(I+M)(I-H)$, but that the distribution is singular and concentrated on the orthogonal complement of the span of the column space of Q. Choosing E:n×(n-p), such that [Q|E] is orthogonal, and noting that $H=QQ^t$ and $I-H=EE^t$, we consider $\mathbf{t}=E^t\mathbf{x}$, which has a nonsingular multivariate Gaussian distribution. Since $\mathbf{r}=E\mathbf{t}$ and $\mathbf{t}=E^t\mathbf{r}$, are 1-1 linearly related, the density of $\mathbf{r}$ is proportional to that of $\mathbf{t}$, which is of the form

$$|V|^{-1/2}\exp(-1/2q)$$

where $V=I+E^tME$ and $q=(\mathbf{t}-E^t\mathbf{m})^tV^{-1}(\mathbf{t}-E^t\mathbf{m})$.

Since M is semi-positive definite, there exists a matrix L, such that $M=LL^t$ and L^tL is positive definite. Noting that in general that $|I+cBB^t|=|I+cB^tB|$, we have that

$$\begin{aligned}|V| &= |I+L^t(EE^t)L| \\ &= |I+L^tL-L^tQQ^tL| \\ &= |I+L^tL|\,|I-Q^tL(I+L^tL)^{-1}L^tQ|.\end{aligned}$$

By the identity (Rao, 1973, p. 73)

$$(A+BDB^t)^{-1}=A^{-1}-A^{-1}B(D^{-1}+B^tA^{-1}B)^{-1}B^tA^{-1} \qquad (A.2)$$

we have that

$$L(I+L^tL)^{-1}L^t=I-(I+M)^{-1} \qquad (A.3)$$

which implies that we can write

$$|V| = |I+M|\,|Q^t(I+M)^{-1}Q| \qquad (A.4)$$

Taking up q, we can write

$$q=(\mathbf{r}-\mathbf{m})^tEV^{-1}E^t(\mathbf{r}-\mathbf{m}),$$

and applying (A.1) to expanding V^{-1} leads to

$$EV^{-1}E^t=(I-H)[I-L\{I+L^t(I-H)L\}^{-1}L^t](I-H).$$

Rewriting $I+L^t(I-H)L$ as $(I+L^tL)-L^tQQ^tL$ and applying (A.2) with $A=I+L^tL$, $B=L^tQ$, and $D=I$, and reducing further by the use of (A.3) and the fact that $(I-H)Q=0$ yields

$$q=\mathbf{d}^t[I+M-Q\{Q^t(I+M)^{-1}Q\}Q^t]\mathbf{d}, \qquad (A.5)$$

where $\mathbf{d}=(I+M)^{-1}(I-H)(\mathbf{r}-\mathbf{m})=(I+M)^{-1}\{\mathbf{r}-(I-H)\mathbf{m}\}$.

Combining (A.4) and (A.5) yields the desired result.

REFERENCES

Barnett, V. and Lewis, T. (1984), *Outliers in Statistical Data*, New York: John Wiley

Beckman, R.J. and Cook, R.D. (1983), "Outlier.....s", *Technometrics*, 25, 119-149.

Cook, R. D. and Weisberg, S. (1980), *Residuals and Influence in Regression*, New York: Chapman and Hall

Hawkins, D.M. (1980), *Identification of Outliers*, London: Chapman and Hall

Lorenzen, T.J and Vance L.C. (1986), "The Economic Design of Control Charts: A Unified Approach," *Technometrics*, 28, 3-10.

Ruppert, D. and Carroll, R.J. (1980), "Trimmed Least Squares Estimates in the Linear Model," *Journal of the American Statistical Association*, 75, 828-838

Tamhane, A.C. (1982), "A Note on the Use of Residuals for Detecting an Outlier in Linear Regression," *Biometrika*, 69, 488-489

Tamhane, A.C and Mah, R.S.H. (1985), "Data Reconciliation and Gross Error Detection in Chemical Process Networks," *Technometrics*, 27, 409-422

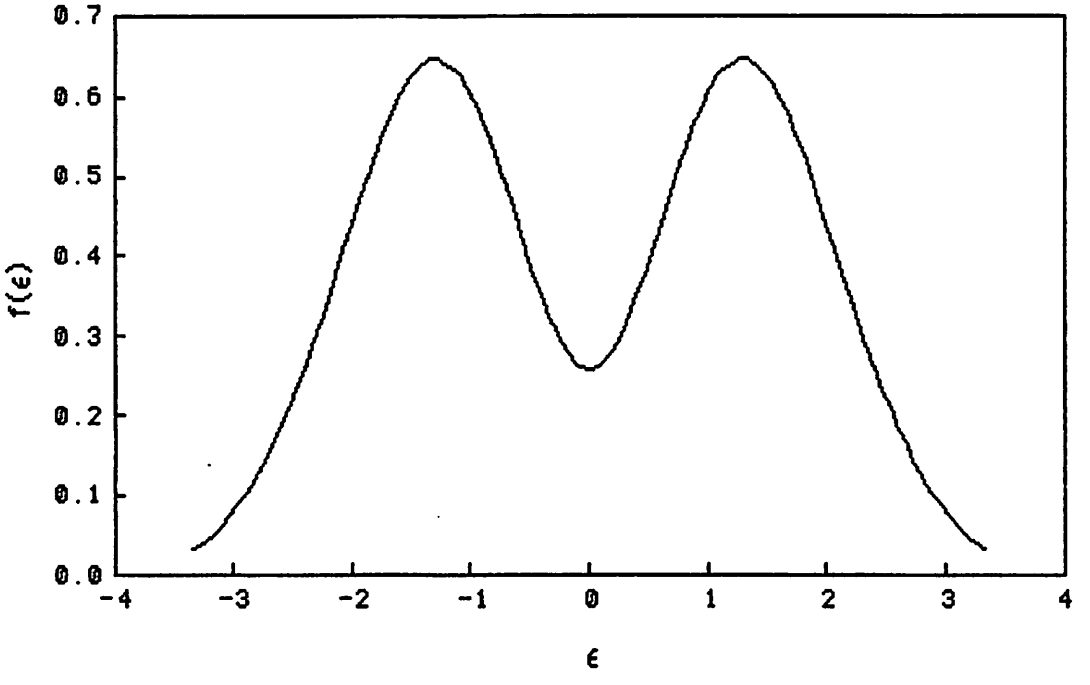

Figure 1. Plot of "contaminating" density in the pure Gaussian model.

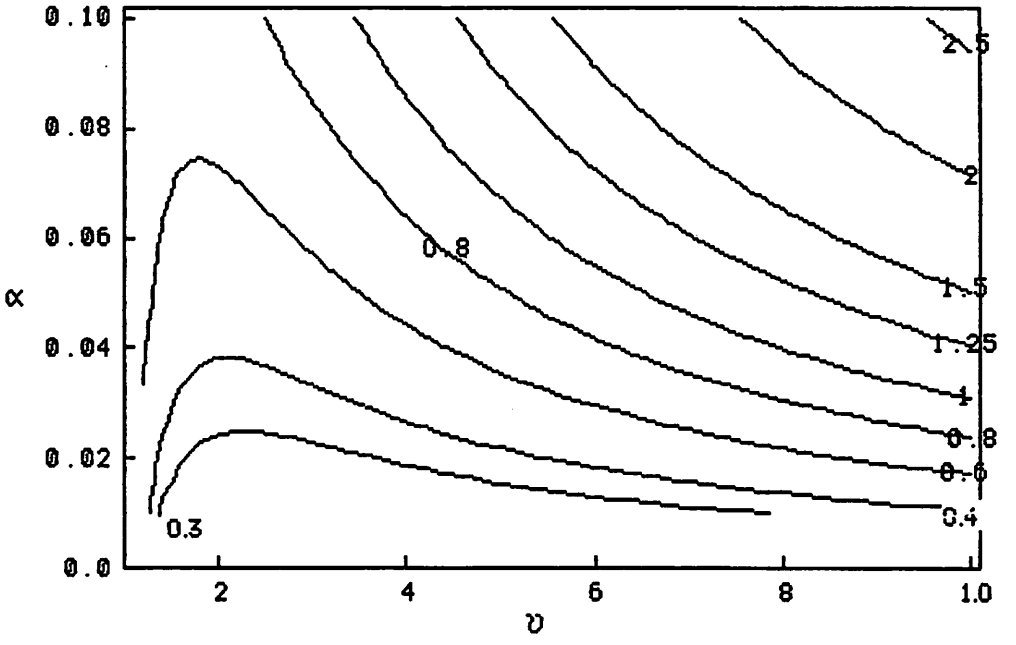

Figure 2. Contours of $h(\alpha,\upsilon)$ for the variance inflated, Gaussian contamination model

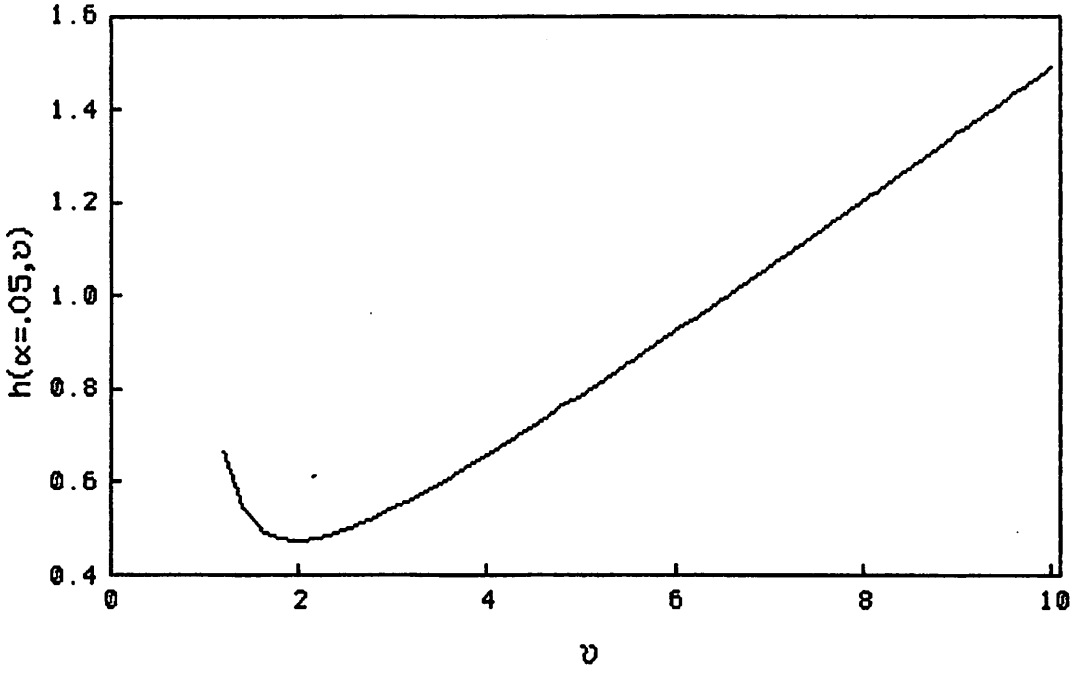

Figure 3. Plot of $h(\alpha,\upsilon)$ for $\alpha=.05$, variance inflation case of the Gaussian contamination model.